stats

MDPI

Article

A Nonparametric Statistical Approach to Content Analysis of Items

Diego Marcondes [1,*,†] and Nilton Rogerio Marcondes [2,†]

1 Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo 05508-090, Brazil
2 Faculdade de Psicologia e de Ciências da Educação, Universidade de Coimbra, 3000-115 Coimbra, Portugal; nilton@fatecpg.com.br
* Correspondence: dmarcondes@ime.usp.br; Tel.: +55-11-97585-6644
† These authors contributed equally to this work.

Received: 6 December 2017; Accepted: 25 January 2018; Published: 1 February 2018

Abstract: In order to use psychometric instruments to assess a multidimensional construct, we may decompose it into dimensions and, in order to assess each dimension, develop a set of items, so one may assess the construct as a whole, by assessing its dimensions. In this scenario, content analysis of items aims to verify if the developed items are assessing the dimension they are supposed to by requesting the judgement of specialists in the studied construct about the dimension that the developed items assess. This paper aims to develop a nonparametric statistical approach based on the Cochran's Q test to analyse the content of items in order to present a practical method to assess the consistency of the content analysis process; this is achieved by the development of a statistical test that seeks to determine if all the specialists have the same capability to judge the items. A simulation study is conducted to check the consistency of the test and it is applied to a real validation process.

Keywords: nonparametric statistics; applied statistics; content validity; psychometric instruments; psychometrics

1. Introduction

Psychometric instruments are built in order to assess psychological constructs that cannot be operationally defined and, consequently, cannot be objectively assessed, such as multidimensional constructs that, according to [1], consist of a number of interrelated attributes or dimensions and exist in multidimensional domains. In order to develop a psychometric instrument to assess a multidimensional construct, a set of items, that assess a dimension, is developed for each one of its dimensions in furtherance of assessing the construct as a whole. The validation process of such an instrument must guarantee that each item assesses its dimension correctly according to desirable characteristics such as reliability and trustworthiness [2].

As psychometric instruments play an important role in researches in the areas of psychology and education, it is necessary that they are thoroughly developed and validated, so that no erroneous results are obtained by their application. The validity of an instrument is divided into four categories: predictive validity, concurrent validity, construct validity and content validity. The first two of these may be considered together as criterion-oriented validation processes [3]. Predictive validity is studied when the instrument assesses a correlated construct to the criterion, providing a prediction for it, and concurrent validity is studied when the instrument is proposed as a substitute for another [3]. The study of construct validity is necessary when the result of the instrument is the measure of an attribute or a characteristic that is not operationally defined, so an instrument is valid when it is possible to determine which construct accounts for the variance of its performance. Furthermore, content validity is established by showing that the instrument items are a sample of a universe in

which the investigator is interested and is ordinarily established deductively, by defining a universe of items and sampling systematically within this universe to build the instrument [3]. Another definition for content validity is that it is the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose [2]. See [4] for more details on instrument validity and [5,6] for practical examples of validation processes.

A list consisting of thirty-five procedures for content validation was proposed by [2]. Amidst these procedures are to match each item to the dimension of the construct that it assesses and request the judgement of specialists in the construct, also called judges, about the developed items. The accomplishment of these procedures is imperative to verify if the developed items are a sample of the universe that the instrument aims to assess. These procedures, components of the theoretical analysis of items, are subjective as they rely on the personal opinions of specialists and researchers. Indeed, the theoretical analysis of items is done by judges and aims to establish the comprehension of the items (semantic analysis) and their pertinence to the attribute that they propose to assess.

This paper aims to propose a nonparametric statistical approach based on the Cochran's Q test to content analysis of items in furtherance of assessing its consistency and reliability. Therefore, our approach does not seek to establish the validity of the instrument, but rather assess the consistency of the content analysis process, so that its rule about the instrument may be trusted. Thus, this approach must be applied among other instrument validation methods, quantitative and qualitative, e.g., semantic analysis, pretrial and factorial analysis, in order to ensure the reliability, consistency, validity and trustworthiness of psychometric instruments.

2. Method

The researcher, supported by the theory of the construct that the instrument aims to assess, develops m items and for each item assigns a theoretical dimension according to the theory and/or his opinion about which dimension the item assesses. Although the items and their dimensions have theoretical foundations, it is necessary to test them in order to determine if every item is indeed assessing the dimension it is supposed to. In order to fulfil such a test, the items are sent to s specialists in the construct, so that they may judge the items according to the dimension they assess. The items may be sent to at least six specialists and should be presented to them in a random order and without their theoretical dimensions, so that their judgement is not biased.

A condition for an item to be excluded from the instrument is determined based on the judgement of the specialists. This condition must exclude the items that do not belong to the universe that the instrument aims to assess, so that the not excluded items are a sample of such a universe. A possible way to proceed is to determine a *Concordance Index* (*CI*) that states that all items in which less than c% of the specialists agree on the dimension they assess must be excluded. One may also take the *Content Validity Ratio* (*CVR*), as proposed by [7], as a condition to exclude items that do not belong to the universe that the instrument aims to assess.

The method to be developed in this paper aims to determine whether all specialists have the same capability to judge the items according to their dimensions, through the analysis of their judgement about the items that were not excluded by the established condition. However, the method does not rank the specialists according to their capabilities, but only determines if all specialists have the same capability, so it is not possible to determine the specialists with low capability.

On the one hand, if there is no evidence that the capabilities of the specialists are different, their judgement is accepted and the items that are not excluded by the established condition are used in the next steps of the instrument validation process. Indeed, if all specialists have the same capability, it may happen that they are all highly capable or almost incapable of judging the items, though the proposed method is not able to differentiate between the two cases. Nevertheless, the two scenarios may be differentiated by a qualitative analysis of the specialists' judgement, by observing if they agree with the theoretical dimension of the items and, when they do not, if there is some theory that supports their choice. Therefore, if their collective judgement is consistent with some theory, then the specialists

may be regarded as all being highly capable of judging the items, given that they all have the same capability to judge them.

On the other hand, if it is determined that the specialists do not all have the same capability to judge the items, then at least one specialist is less capable to judge them than the others, which may bias the validation of the instrument. Therefore, in such a scenario, we propose two approaches in order to avoid a biased validation process. First, we propose that the specialists' judgement be disregarded and a new group of specialists be requested to judge the items. However, this approach may be impractical, as time and resources may be too limited to repeat the cycle of specialists' judgement more than once. Nonetheless, we propose a much more practical approach that consists in applying the proposed method to all subgroups of specialists of size s^*, $6 \leq s^* < s$, of the original group of specialists, and then choosing the judgement of the subgroup whose specialists all have the same capability to judge the items. This approach is presented in more detail in the application section.

3. Notation and Definitions

Let $C = \{C_1, \ldots, C_n\}$ be a construct divided in n dimensions and U be the universe of all the items that assess the dimensions of C. A set $I = \{i_1, \ldots, i_m\}$ of m items is developed based on the theory about C and then a subset $I^* \subset I$ of items, that we believe to be a subset of U, is determined, by the following process.

Denote $E = \{e_1, \ldots, e_s\}$ a set of s specialists and let $C_{c(i_l)} \in C$ be the dimension that the item $i_l \in I^*$ assesses. Let the random variables $\{X_{i_l}(e_j) : i_l \in I, e_j \in E\}$, defined on $(\Omega, \mathbb{F}, \mathbb{P})$, be so that $X_{i_l}(e_j) = k$ if the specialist e_j judged the item i_l at the *kth* dimension of C (in the following, we have that l goes from 1 to m and that j goes from 1 to s).Note that if $i_l \in I^*$ and $X_{i_l}(e_j) = c(i_l)$, then the specialist e_j judged the item i_l correctly. The capability of the specialist e_j to judge the items is defined as

$$\boldsymbol{P}(e_j) = (P_{i_l}(e_j) : i_l \in I^*)$$

in which $P_{i_l}(e_j) = \mathbb{P}\{X_{i_l}(e_j) = c(i_l)\}, \forall i_l \in I^*$ and $\forall e_j \in E$. In the proposed approach, denoting $|I^*|$ the length of I^*, we are interested in developing a hypothesis test to determine if $\boldsymbol{P}(e_j) = \boldsymbol{p} \in [0,1]^{|I^*|}, \forall e_j \in E$, i.e., if all specialists have the same capability to judge the items.

For this purpose, let a random sample of the judgement of the specialist e_j about the items of I be given by $\boldsymbol{x}_{e_j} = \{x_{i_1}(e_j), \ldots, x_{i_m}(e_j)\}$ and let $\mathbb{X}$ be the space of all possible random samples $\{\boldsymbol{x}_{e_j} : e_j \in E\}$. Define the random sets $\{M_{i_l} : i_l \in I\}$ as

$$M_{i_l} = \arg \max_{k \in \{1,\ldots,n\}} \left\{ \sum_{e_j \in E} \mathbb{1}_{\{k\}} \left(X_{i_l}(e_j) \right) \right\}$$

in which $\mathbb{1}_{\{A\}}(\cdot)$ is the indicator function of the set A. Note that M_{i_l} is the set containing the number of the dimensions in which the majority of the specialists judged the item $i_l \in I$. Given a random sample $\{\boldsymbol{x}_{e_j} : e_j \in E\} \in \mathbb{X}$ and a subset $I^* \subset I$ of items, the set $\{m_{i_l} : i_l \in I^*\}$, determined from the sample values $\{\boldsymbol{x}_{e_j} : e_j \in E\}$, is a random sample of $\{M_{i_l} : i_l \in I^*\}$.

The subset I^* may be defined by a condition function, a function of the sample $\{\boldsymbol{x}_{e_j} : e_j \in E\}$, given by $f : \mathbb{X} \mapsto \mathcal{P}(I)$, in which $\mathcal{P}(\cdot)$ is the power set operator. The condition function must be such that if $\{m_{i_l} : i_l \in I^*\}$ is determined from $\{\boldsymbol{x}_{e_j} : e_j \in E\} \in \mathbb{X}$ and $I^* = f(\boldsymbol{x}_{e_j} : e_j \in E)$, then the length of m_{i_l} is one, $\forall i_l \in I^*$. The *CI* for $c > 50$ and the *CVR* are condition functions. As the *CI* may be obtained from other concordance indexes, as the *Content Validity Index* (*CVI*) that is used to measure concordance when the construct is unidimensional and the task of the specialists is to judge the item's relevance [8–11], the method developed in this paper may also be applied in other scenarios. From now on, it is supposed that the condition function may be expressed as a *CI*.

The condition function is based on the assumption that an item is in the universe of items that assess the construct of interest if the majority of specialists agree on the dimension it assesses. Of course,

one may take a different criterion to exclude the items that do not assess the construct of interest, although our method may be applied only if the criterion can be expressed as a condition function, as it is based on the fact that M_{i_l} is a univariate random variable.

Finally, define

$$W_{i_l}(e_j) = \mathbb{1}_{\{M_{i_l}\}}\Big(X_{i_l}(e_j)\Big)$$

as the random variable that indicates if the specialist e_j judged the item i_l at the same dimension as the majority of the specialists. Given a random sample $\{\boldsymbol{x}_{e_j} : e_j \in E\} \in \mathbb{X}$ and a subset $f(\boldsymbol{x}_{e_j} : e_j \in E) = I^* \subset I$ of items, the set $\{w_{i_l}(e_j) : i_l \in I^*, e_j \in E\}$, determined from the sample values $\{\boldsymbol{x}_{e_j} : e_j \in E\}$, is a random sample of $\{W_{i_l}(e_j) : i_l \in I^*, e_j \in E\}$.

On the one hand, whilst we observe the values of the random variables $\{X_{i_l}(e_j) : i_l \in I, e_j \in E\}$, we do not know if the specialists judged the items correctly or not, as the dimension that an item really assesses (if any) is unknown. Therefore, it is not possible to differentiate the specialists by the number of items they judged correctly, for example. On the other hand, from the random variables $\{W_{i_l}(e_j) : i_l \in I, e_j \in E\}$, we know the concordance of the specialists on the judgement of the items, which gives us a relative measure of their capability to judge the items. Therefore, we are able to test if all the specialists have the same capability to judge the items by applying the Cochran's Q test, although we cannot determine the capability of each one.

4. Assumptions

The development of the items and the judgement of the specialists must satisfy two assumptions so that the method presented below may be applied:

1. Each item $i_l \in I^*$ assesses one, and only one, dimension $C_{c(i_l)} \in C$.
2. The random variables $\{X_{i_l}(e_j) : i_l \in I^*, e_j \in E\}$ are independent.

Assumption **1** establishes that the items that were not excluded by the condition function, i.e., the items in I^*, are well constructed and assess only one dimension of C, while Assumption **2** imposes that the specialists judge the items independently of each other and that the judgement of a specialist about one item does not depend on his judgement about any other item. These assumptions are not strong, as it is expected that they are satisfied if the items were well constructed. Indeed, the better the condition function in determining what items are not in U, the better the quality of the items in I^*. Therefore, the assumptions above are closely related to the condition function. If, in fact, $I^* \subset U$, then the first assumption is immediately satisfied, as there is no intersection between two dimensions of a construct, and the second assumption may also hold, as the items are well defined.

5. Mathematical Deduction

Given a random sample $\{\boldsymbol{x}_{e_j} : e_j \in E\} \in \mathbb{X}$, it is not trivial to estimate the capabilities $\{\boldsymbol{P}(e_j) : e_j \in E\}$, as the dimension that each item assesses is unknown. Examining such a random sample, it is known that the specialist e_j judged the item i_l at the dimension C_k, but it is not possible to determine, with probability 1, if he judged such an item correctly. Therefore, the problem is, given a random sample $\{\boldsymbol{x}_{e_j} : e_j \in E\} \in \mathbb{X}$, to determine random variables that allow us to test if the capability of all the specialists is the same. It will be shown that if the random variables $\{W_{i_l}(e_j) : e_j \in E\}$ are not identically distributed $\forall i_l \in I^*$, then the specialists do not all have the same capability to judge the items. Indeed, in order to test if the capability of all specialists is the same, we consider the following null hypotheses:

$$H_0 : \begin{cases} \textbf{1.} & \boldsymbol{P}(e_j) = (p^{(i_1)}, \ldots, p^{(i_{|I^*|})}) = \boldsymbol{p} \in [0,1]^{|I^*|}, \forall e_j \in E \\ \textbf{2.} & \Big(\mathbb{P}\{X_{i_l}(e_j) = 1\}, \ldots, \mathbb{P}\{X_{i_l}(e_j) = n\}\Big) \text{ is a permutation of} \\ & \Big(p^{(i_l)}, p_1^{(i_l)}, \ldots, p_{n-1}^{(i_l)}\Big) \in [0,1]^n, p^{(i_l)} + \sum_{k=1}^{n-1} p_k^{(i_l)} = 1, \forall i_l \in I^*, \forall e_j \in E \end{cases}$$

Of course, we are only interested in testing the first part of H_0, that refers to the capability of the specialists, i.e., that all specialists have the same capability to judge the items. However, the second part is needed to develop a test statistic for H_0. It will be argued that for great values of $p^{(i_l)}$, the hypothesis that is actually being tested is the first one.

The propositions below set the scenario for the nonparametric test, i.e., the Cochran's Q test, that is used to test H_0.

Proposition 1. *The random variables* $\{W_{i_l}(e_j) : i_l \in I^*\}$ *are independent* $\forall e_j \in E$, *but the random variables* $\{W_{i_l}(e_j) : e_j \in E\}$ *are dependent* $\forall i_l \in I^*$.

Proof. On the one hand, the random variables $\{W_{i_l}(e_j) : i_l \in I^*\}$ are each, by assumption **2**, function of independent random variables, therefore they are independent. On the other hand, note that $\sum_{e_j \in E} W_{i_l}(e_j) \geq \lceil \frac{cs}{100} \rceil$, for at least c% of the specialists must agree on the dimension that an item in I^* assesses, which establishes a dependence. □

Proposition 2. *Under* H_0, *the random variables* $\{W_{i_l}(e_j) : e_j \in E\}$ *are identically distributed for all* $i_l \in I^*$.

Proof. We have that

$$\begin{aligned}\mathbb{P}\{W_{i_l}(e_j) = 1\} &= \mathbb{P}\{X_{i_l}(e_j) = M_{i_l}\} \\ &= \mathbb{P}\{X_{i_l}(e_j) = c(i_l), M_{i_l} = c(i_l)\} + \mathbb{P}\{X_{i_l}(e_j) = M_{i_l}, M_{i_l} \neq c(i_l)\}.\end{aligned}$$

Now let $X^{(i_l)} \sim Binomial(s-1, p^{(i_l)})$ and $X_k^{(i_l)} \sim Binomial(s-1, p_k^{(i_l)}), k \in \{1, \ldots, n-1\}$, be independent random variables, and let $f^* = \lfloor (\frac{c}{100})s \rfloor$, in which c is the *CI*. Then, under H_0,

$$\begin{aligned}\mathbb{P}\{X_{i_l}(e_j) = c(i_l), M_{i_l} = c(i_l)\} &= \mathbb{P}\{M_{i_l} = c(i_l) | X_{i_l}(e_j) = c(i_l)\}\mathbb{P}\{X_{i_l}(e_j) = c(i_l)\} \\ &= \mathbb{P}\{X^{(i_l)} \geq f^*\}p^{(i_l)}\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\{X_{i_l}(e_j) = M_{i_l}, M_{i_l} \neq c(i_l)\} &= \sum_{\substack{k=1 \\ k \neq c(i_l)}}^{n} \mathbb{P}\{X_{i_l}(e_j) = k, M_{i_l} = k\} \\ &= \sum_{\substack{k=1 \\ k \neq c(i_l)}}^{n} \mathbb{P}\{M_{i_l} = k | X_{i_l}(e_j) = k\}\mathbb{P}\{X_{i_l}(e_j) = k\} \\ &= \sum_{k=1}^{n-1} \mathbb{P}\{X_k^{(i_l)} \geq f^*\}p_k^{(i_l)}.\end{aligned}$$

Hence,

$$\mathbb{P}\{W_{i_l}(e_j) = 1\} = \mathbb{P}\{X^{(i_l)} \geq f^*\}p^{(i_l)} + \sum_{k=1}^{n-1} \mathbb{P}\{X_k^{(i_l)} \geq f^*\}p_k^{(i_l)}$$

which does not depend on e_j and the result follows. □

It is important to note that if all $p^{(i_l)}$ are approximately 1, then $\mathbb{P}\{W_{i_l}(e_j) = 1\} \approx p^{(i_l)}\mathbb{P}\{X \geq f^*\}$ and the hypothesis that is actually being tested is the first part of H_0. Therefore, it is reasonable to test H_0 in order to determine if all the specialists have the same capability to judge the items, as, if it is indeed true, we expect that all $p^{(i_l)}$ are great and the second part of H_0 will hardly lead to the rejection of H_0 when the capability is the same.

This test may be used as a diagnostic for the content analysis of items. If H_0 is not rejected, then there is no evidence that the capabilities of the specialists are different. However, if H_0 is rejected, we do not know if it is the first or the second part (or both) of H_0 that is not being satisfied by the judgement of the specialists. Nevertheless, we may disregard their judgement in any case, as either their capability is not the same or they are the same, but some $p^{(i_l)}$ are small, which led to the rejection of H_0 by its second part.

6. Hypothesis Testing

The Cochran's Q test may be applied to the random sample $\{w_{i_l}(e_j) : i_l \in I^*, e_j \in E\}$ determined from $\{\boldsymbol{x}_{e_j} : e_j \in E\}$ as a way to test H_0 [12]. The assumptions of the Cochran's Q test, using the notation of this paper, are as follows:

(a) The items of I^* were randomly selected from the items that form the universe U that the instrument aims to assess.
(b) The random variables $\{W_{i_l}(e_j) : i_l \in I^*, e_j \in E\}$ are dichotomous.
(c) The random variables $\{W_{i_l}(e_j) : i_l \in I^*\}$ are independent.

The Cochran's Q test is used in applications in which treatments are applied independently to blocks (subjects) and the result of each treatment application is either a success or a failure (zero or one) [13]. In our case, we have that the items may be seen as the *blocks* and the specialists as the *treatments*. What the Cochran's Q test evaluates is if the treatments are all equally effective or, in our case, if the specialists are all equally capable of judging the items (which is equivalent to testing if the random variables $\{W_{i_l}(e_j) : e_j \in E\}$ are identically distributed for all $i_l \in I^*$). Therefore, if we reject the null hypothesis, we conclude that $\{W_{i_l}(e_j) : e_j \in E\}$ are not identically distributed for all $i_l \in I^*$ and, by Proposition 2, H_0 is also rejected. Thus, the hypothesis tested by the Cochran's Q test is indeed H_0.

The statistic of the test is calculated from Table 1, in which $I^* = \{i_1^*, \ldots, i_v^*\}$, and may be expressed as

$$Q = \sum_{r=1}^{s} \frac{s(s-1)\left(D_r - \frac{N}{s}\right)^2}{\sum_{l=1}^{v} R_l(s - R_l)}$$

Table 1. Table of the observed random sample.

Item	Specialist			Total
	e_1	$\cdots$	e_s	
i_1^*	$w_{i_1^*}(e_1)$	$\cdots$	$w_{i_1^*}(e_s)$	$R_1 = \sum_{e_j \in E} w_{i_1^*}(e_j)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i_v^*	$w_{i_v^*}(e_1)$	$\cdots$	$w_{i_v^*}(e_s)$	$R_v = \sum_{e_j \in E} w_{i_v^*}(e_j)$
Total	$D_1 = \sum_{i_l \in I^*} w_{i_l}(e_1)$	$\cdots$	$D_s = \sum_{i_l \in I^*} w_{i_l}(e_s)$	$N = \sum_{i_l \in I^*} \sum_{e_j \in E} w_{i_l}(e_j)$

The exact distribution of the Q statistic may be calculated by the method presented by [14], although a large sample approximation may be used instead. If $|I^*|$ is large, then the distribution of Q is approximately χ^2 with $(s-1)$ degrees of freedom [13].

It is worth mentioning that the random variables $\{W_{i_l}(e_j) : e_j \in E\}$ being identically distributed for all $i_l \in I^*$ does not imply that all the specialists have the same capability to judge the items, although there is no evidence that their capabilities are different. If there is no evidence that the capabilities of the specialists to judge the items are different, their judgement may be accepted.

If it is determined that the random variables $\{W_{i_l}(e_j) : e_j \in E\}$ are not identically distributed for all $i_l \in I^*$, then the judgement of the specialists is disregarded as H_0 is rejected. The items may be

judged by different groups of specialists until they are judged by one in which all the specialists have the same capability to judge the items. These groups may be formed by new specialists or may be a subgroup of size s^*, $6 \leq s^* < s$, of the specialists for which H_0 was rejected.

7. Simulation Study

As the Cochran's Q test is not a powerful one, i.e., its Type I error may be too great, a simulation study is conducted to estimate its power in some specific cases. The power of a statistical test is defined as the probability of H_0 being rejected when it is false and depends on the real scenario, i.e., on the real values of the parameters considered on H_0. Therefore, the power of the Cochran's Q test in testing H_0 depends on the real capability of each specialist to judge the items, so the simulation study considers 10 distinct scenarios and is conducted as follows.

For each scenario, we simulate 50,000 judgements of the same items by the specialists and then determine the proportion of the simulations in which H_0 was rejected at a significance, i.e., Type II error, of 5%. This proportion is regarded as an estimate for the power of the test in the considered scenario. A CI of 50% is used to determine I^* in each simulation. The results of all 10 scenarios provide a wide picture of the power of the test, so we will know for which scenarios it is more powerful.

We consider in all scenarios nine specialists judging 30 items into three dimensions; this is the framework of the application in the next section. We also consider that the capability of each specialist is the same for all items, i.e., that $\mathbb{P}\{X_{i_l}(e_j) = c(i_l)\} = p_j$ for all $j \in \{1, \ldots, 9\}$ and $l \in \{1, \ldots, |I^*|\}$. Finally, we assume that $\mathbb{P}\{X_{i_l}(e_j) = k\} = (1 - p_j)/2$ for all $k \neq c(i_l)$, $j \in \{1, \ldots, 9\}$ and $l \in \{1, \ldots, |I^*|\}$. A pseudocode for the simulation of each scenario is presented in Algorithm 1. The scenarios and their estimated test power are displayed in Table 2.

Algorithm 1 Pseudocode that estimates the power of the Cochran's Q test under a given scenario from 50,000 simulated judgements.

Ensure: rejected $= 0$
for simulation $\in$ {1,...,50,000} **do**
  for $j \in \{1, \ldots, 9\}$ **do**
    for $l \in \{1, \ldots, 30\}$ **do**
      Simulate $X_{i_l}(e_j)$ from the Multinomial with parameter $\left(p_j, \frac{(1-p_j)}{2}, \frac{(1-p_j)}{2}\right)^*$
    end for
  end for
  Determine I^* as the items such that at least 5 specialists agree on the dimension they assess
  Calculate $\{M_{i_l} : i_l \in I^*\}$
  Calculate $\{W_{i_l}(e_j) : i_l \in I^*, j \in \{1, \ldots, 9\}\}$
  Calculate the statistic Q
  Calculate the p-value of Q
  if p-value < 0.05 **then**
    rejected $=$ rejected $+ 1$
  end if
end for
return rejected/50,000

* In scenarios 1 to 8. In scenarios 9 and 10 the Multinomial has parameter $(p_j, \xi_j, 1 - p_j - \xi_j)$ in which ξ_j is simulated from a uniform distribution with range $[0, 1 - p_j]$.

Table 2. The estimated power of the test for each scenario.

Scenario	Description	Items *	Power
1	$p_j = 0.9, j \neq 1$ and $p_1 = 0.45$	29.9	0.9936
2	$p_j = 0.9, j \notin \{1,2,3\}$ and $p_1 = p_2 = p_3 = 0.45$	29.2	0.9999
3	$p_j = 0.9, j \notin \{1,2,3\}$ and $p_1 = 0.45, p_2 = 0.35, p_3 = 0.25$	29.7	1
4	$p_j = 0.9, j \neq 1$ and $p_1 = 0.8$	29.9	0.1601
5	$p_j = 0.9, j \notin \{1,2\}$ and $p_1 = p_2 = 0.8$	29.9	0.2421
6	$p_j = 0.6, j \notin \{1,2,3\}$ and $p_1 = p_2 = p_3 = 0.75$	25.5	0.2342
7	$p_j = 0.3, j \notin \{1,2,3\}$ and $p_1 = p_2 = p_3 = 0.75$	14.7	0.2778
8	$(p_1, \ldots, p_9) = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$	17.3	0.9001
9	$p_j = 0.9, \forall j$, but the second part of H_0 is not true	29.9	0.0454
10	$p_j = 0.6, \forall j$, but the second part of H_0 is not true	23.1	0.0469

* The mean number of items not excluded by the *CI*.

On the one hand, we see in Table 2, that the power of the test is great when the majority of the specialists have the same high capability, while few specialists have a low capability, as is the case for scenarios 1, 2 and 3. This is also the case for scenario 8, when the specialists have different capability and there are specialists whose capability is very low. On the other hand, the power of the test is quite low when some of the specialists have the same high capability, and the specialists with lower capability are almost as capable as them, as is the case for scenarios 4, 5 and 6.

In scenario 7, we see that the power of the test is low when the majority of specialists have the same low capability (0.3 is this case). It happens because the specialists hardly agree on the dimension that each item assesses (as some of them are not capable) so many items are excluded by the *CI* and, for the items that remain, the not capable specialists agree with the highly capable ones, so it seems that they have high capability. Indeed, in scenario 7, the mean number of not excluded items is the lowest of all scenarios, so a low concordance among the specialists is evidence of the existence of specialists of low capability, given that the items were well constructed.

Finally, as pointed out in the Mathematical Deduction section, we see in scenarios 9 and 10 that the hypotheses that is actually being tested when all the specialists are highly and equally capable is the first part of H_0, as the power of the test is close to the Type II error, which must be the case if the hypothesis is true.

The simulation study shed light on some interesting facts about the proposed method in the considered scenarios. On the one hand, if the majority of the specialists have a homogeneous high capability, and few specialists have a very low capability, or if the capability of the specialists is highly heterogeneous, then the power of the test is great. However, if the specialists all have high, but different, capability then the power of the test is low. On the other hand, if the majority of the specialists have a low capability, then a great number of items is excluded by the *CI* and, given that the items were well constructed, we may conclude that the specialists have low capability of judging the items, even though the power of the test is low. Finally, if only the first part of H_0 is being satisfied, and the capability of the specialists is high, then the power of the test is low and, therefore, the hypothesis that is really being tested is the first part of H_0.

8. Application: Perception about the Evaluation of the Teaching-Learning

In this section, we apply the developed method to a real validation process, in order to analyse the content of items of an instrument that aims to assess the perception of teachers and students of higher education institutions about the teaching-learning process; this is a construct that may be divided into three dimensions: process (P), judgement (J) and teaching-learning (T).

The evaluation of teaching-learning is a process, as it must have a well defined beginning, middle and end and must have a continuous, cumulative and systematic character. Indeed, it is a systematic mechanism for gathering information over time, with well defined levels, which characterises it as a process. Also, the evaluation of teaching-learning has a judgement dimension because it must issue a

judgement of value or assign a score through the analysis of educational results obtained from the information gathered over time. Finally, the evaluation of teaching-learning has a teaching-learning dimension because, as indicated by its own name, it must not only evaluate the learning, but also the teaching: it should not only evaluate what the student has learnt, but also what the teacher has taught. Therefore, the evaluation of teaching-learning is a process of data gathering, in which an individual judges or is judged according to the teaching-learning.

In order to develop an instrument to assess this construct, 30 items were developed and sent to nine specialists; they would judge the items according to the dimension that, in their opinion, each one assesses. The condition defined for excluding an item is the CI with $c = 50$. The judgements of the specialists are presented in Table 3; the table for the Cochran's Q test is displayed in Table 4 and a translation of the items, that were originally constructed in Portuguese, is presented in the Appendix A.

Table 3. Judgement of the specialists about each item, i.e., the sample $\{\boldsymbol{x}_{e_j} : e_j \in E\}$.

Item	Specialist									Dimension *	Theoretical Dimension
	1	2	3	4	5	6	7	8	9		
1	T	P	P	T	T	T	T	P	P	T	P
2	P	J	T	T	P	P	T	P	T	-	T
3	T	T	J	P	P	P	P	J	J	-	J
4	J	P	P	P	P	P	P	J	J	P	P
5	T	J	T	T	T	T	J	T	P	T	P
6	P	T	T	P	T	T	T	P	T	T	P
7	J	J	J	J	J	J	J	J	J	J	J
8	T	P	P	T	P	P	T	P	T	P	P
9	J	P	T	J	T	T	P	P	P	-	T
10	J	J	J	J	J	J	J	J	J	J	J
11	J	T	J	J	J	J	J	J	J	J	J
12	P	T	T	P	P	P	P	J	P	P	P
13	J	P	P	T	J	J	J	J	T	J	J
14	T	T	T	T	P	P	T	P	T	T	P
15	J	P	J	J	J	J	J	J	J	J	J
16	T	P	T	T	P	P	J	P	T	-	P
17	P	P	P	P	T	T	P	T	P	P	T
18	J	T	T	T	P	P	J	T	J	-	T
19	T	T	T	T	P	P	T	J	P	T	T
20	P	P	P	T	P	P	P	J	J	P	P
21	J	J	J	J	J	J	J	J	P	J	P
22	P	P	P	P	P	P	T	T	P	P	P
23	J	J	J	J	J	J	J	J	J	J	J
24	T	J	P	T	P	P	T	J	T	-	J
25	J	J	J	J	J	J	J	J	J	J	J
26	T	P	T	T	P	P	P	P	T	P	T
27	P	P	P	T	P	P	T	J	P	P	T
28	T	P	P	J	P	P	J	P	T	P	T
29	T	T	P	T	P	P	P	P	T	P	T
30	T	J	J	J	T	T	J	J	P	J	J

* The dimension on which at least 50% of the specialists agree that the item assesses.

Table 4. Table for the Cochran's Q test, i.e., the sample $\{w_{i_l}(e_j) : i_l \in I^*, e_j \in E\}$.

Item	Specialist									Total
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
1	1	0	0	1	1	1	1	0	0	5
4	0	1	1	1	1	1	1	0	0	6
5	1	0	1	1	1	1	0	1	0	6
6	0	1	1	0	1	1	1	0	1	6
7	1	1	1	1	1	1	1	1	1	9
8	0	1	1	0	1	1	0	1	0	5
10	1	1	1	1	1	1	1	1	1	9
11	1	0	1	1	1	1	1	1	1	8
12	1	0	0	1	1	1	1	0	1	6
13	1	0	0	0	1	1	1	1	0	5
14	1	1	1	1	0	0	1	0	1	6
15	1	0	1	1	1	1	1	1	1	8
17	1	1	1	1	0	0	1	0	1	6
19	1	1	1	1	0	0	1	0	0	5
20	1	1	1	0	1	1	1	0	0	6
21	1	1	1	1	1	1	1	1	0	8
22	1	1	1	1	1	1	0	0	1	7
23	1	1	1	1	1	1	1	1	1	9
25	1	1	1	1	1	1	1	1	1	9
26	0	1	0	0	1	1	1	1	0	5
27	1	1	1	0	1	1	0	0	1	6
28	0	1	1	0	1	1	0	1	0	5
29	0	0	1	0	1	1	1	1	0	5
30	0	1	1	1	0	0	1	1	0	5
Total	17	17	20	16	20	20	19	14	12	155

The statistic of the Cochran's Q test for the data in Table 4 is $Q = 13.8$ and the test p-value is 0.087, so there is no evidence that H_0 is not true, at a significance of 5%. Furthermore, as the majority of the specialists agreed on the dimension that 24 out of 30 (80%) items assess, we also do not have evidence that the capability of the specialists is low. Therefore, based on the proposed method, there is no reason to disregard the judgement of the specialists.

Nevertheless, in order to illustrate the proposed approach for the case in which H_0 is rejected, we apply the test to every subgroup of size $6 \leq s^* < 9$ of specialists, which amounts to 130 subgroups, and see for which subgroups the capability of the specialists is the same. From the 130 subgroups, for 29 of them H_0 was rejected at a significance of 5%. The Q statistic and the p-value for the 10 groups with the greatest p-values are displayed in Table 5. If H_0 had been rejected by a group of nine specialists, we could then look for a subgroup of these specialists for which H_0 is not rejected and, with the help of a qualitative analysis, we could choose a subgroup of these specialists instead of disregarding their judgement as a whole and sending the items to other specialists to judge.

Table 5. Result of the Cochran's Q test for the subgroups of specialists with the greatest p-values.

Specialists	Q	p-Value
(1,3,4,5,6,7)	1.064	0.957
(1,2,3,4,5,7)	1.818	0.874
(1,2,3,4,6,7)	1.818	0.874
(1,2,3,4,5,6)	2.340	0.800
(1,2,3,5,7,9)	2.391	0.793
(1,2,3,6,7,9)	2.391	0.793
(2,3,4,5,6,7)	2.519	0.774
(1,2,4,5,6,9)	2.528	0.772
(1,3,4,5,6,9)	2.619	0.758
(1,3,4,5,6,7,9)	3.612	0.729

9. Final Remarks

The Cochran's Q test is not a powerful one, so the method must be used with caution. The validation of a psychometric instrument is a process that comprises various procedures, therefore it must not be restricted to content analysis of items and the method developed in this paper. It is important to apply other validation techniques, both qualitative and quantitative, to the instrument to properly validate it.

The method may be improved in order to further decrease the subjectivity of the content analysis of items, especially by the development of more powerful tests and the definition of other random variables that enable the comparison between the judgement of the specialists. This paper does not exhaust the subject, but presents a nonparametric statistical approach that aims to decrease the subjectivity of a subjective process and that may applied not only to content analysis of items, but also to any statistical application that enables the definition of variables such as those of this paper.

Supplementary Materials: The **R** (A Language and Environment for Statistical Computing) script used in the simulation study and in the application section is available online at http://www.mdpi.com/2571-905X/1/1/1/s1.

Acknowledgments: We would like to thank Eduardo João Ribeiro dos Santos, Joaquim Armando Gomes Alves Ferreira and Maria Cristina Pereira Matos for the orientation on the Ph.D. thesis on which the instrument used in Section 8 was developed.

Author Contributions: D.M. wrote the paper and developed the statistical deduction and the simulation studies. N.R.M. developed the instrument and collected the data used in Section 8.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CI	Concordance Index
CVR	Content Validity Ratio
CVI	Content Validity Index
P	Process
J	Judgement
T	Teaching-Learning

Appendix A. Items Developed in the Application Section

A translation of the constructed items, whose response is the Likert Scale, and their theoretical dimension, are presented below. The items were originally written in Portuguese.

1. The evaluation is an instrument strategically used to help in the difficulties (Process).
2. The evaluation assumes a formative role in the teaching-learning process (Teaching-Learning).
3. The proposed evaluation methods are fair and appropriate (Judgement).
4. The time available for evaluation is sufficient (Process).
5. The evaluation offers recovery strategies for students that have difficulties (Process).
6. The instructions given for the assignments subjected to evaluation are useful (Process).
7. The evaluation is a tool for punishing the student in some manner (Judgement).
8. The evaluation is an essential tool for the teaching-learning process (Process).
9. The evaluation is an essential tool for the understanding of the taught subject (Teaching-Learning).
10. The evaluation is a process that ranks the students in some manner (Judgement).
11. The evaluation is a process that, in a particular way, builds a hierarchy among the students (Judgement).
12. The evaluation is a process that follows the student during all his academic life (Process).
13. The evaluation has different meanings for who evaluate and for who is evaluated (Judgement).

14. The evaluation is used to find out where and how the teaching-learning may be improved (Process).
15. The evaluation is a tool to reward the student in some manner (Judgement).
16. The evaluation is a tool to diagnostic the teaching-learning process (Process).
17. The evaluation is a tool with technical and pedagogical characteristics (Teaching-Learning).
18. The evaluation aims to identify how much the student has learnt in the subjects (Teaching-Learning).
19. The evaluation aims to identify which paths to take to obtain knowledge (Teaching-Learning).
20. The evaluation is a systematic evidence gathering process (Process).
21. The evaluation is a process of outlining, obtaining and providing information that permits to judge decision alternatives (Process).
22. The evaluation is a process with continuous, cumulative and systematic, but not episodic, character (Process).
23. Evaluate means to provide a judgement of value or to assign a score to whom is being evaluated (Judgement).
24. The evaluation is a tool that permits to inquire to what extent the defined objectives are being achieved (Judgement).
25. The evaluation has an authoritarian and classificatory role inside the process of teaching-learning (Judgement).
26. The evaluation is an educational component that can facilitate the teaching-learning process (Teaching-Learning).
27. The teaching-learning process and the evaluation are not isolated parts of the education process (Teaching-Learning).
28. The evaluation identifies the more adequate path to make excellent teaching-learning feasible (Teaching-Learning).
29. The evaluation stimulates the acts of teaching and learning as a simultaneous process (Teaching-Learning).
30. The evaluation involves the intentional judgement of a process developed by an individual, during his learning (Judgement).

References

1. Law, K.S.; Wong, C.S.; Mobley, W.M. Toward a taxonomy of multidimensional constructs. *Acad. Manag. Rev.* **1998**, *23*, 741–755.
2. Haynes, S.N.; Richard, D.; Kubany, E.S. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychol. Assess.* **1995**, *7*, 238–247.
3. Cronbach, L.J.; Meehl, P.E. Construct validity in psychological tests. *Psychol. Bull.* **1955**, *52*, 281–302.
4. Cook, D.A.; Beckman, T.J. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am. J. Med.* **2006**, *119*, doi:10.1016/j.amjmed.2005.10.036.
5. Aladwani, A.M.; Palvia, P.C. Developing and validating an instrument for measuring user-perceived web quality. *Inf. Manag.* **2002**, *39*, 467–476.
6. Engel, S.G.; Wittrock, D.A.; Crosby, R.D.; Wonderlich, S.A.; Mitchell, J.E.; Kolotkin, R.L. Development and psychometric validation of an eating disorder-specific health-related quality of life instrument. *Int. J. Eat. Disord.* **2006**, *39*, 62–71.
7. Lawshe, C.H. A quantitative approach to content validity. *Pers. Psychol.* **1975**, *28*, 563–575.
8. Wynd, C.A.; Schmidt, B.; Schaefer, M.A. Two quantitative approaches for estimating content validity. *West. J. Nurs. Res.* **2003**, *25*, 508–518.
9. Rubio, D.M.; Berg-Weger, M.; Tebb, S.S.; Lee, E.S.; Rauch, S. Objectifying content validity: Conducting a content validity study in social work research. *Soc. Work Res.* **2003**, *27*, 94–104.

10. Polit, D.F.; Beck, C.T. The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Res. Nurs. Health* **2006**, *29*, 489–497.
11. Polit, D.F.; Beck, C.T.; Owen, S.V. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res. Nurs. Health* **2007**, *30*, 459–467.
12. Cochran, W.G. The comparison of percentages in matched samples. *Biometrika* **1950**, *37*, 256–266.
13. Conover, W. *Practical Nonparametric Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 1998.
14. Patil, K.D. Cochran's Q test: Exact distribution. *J. Am. Stat. Assoc.* **1975**, *70*, 186–189.